Using Gaze Data to Predict Multiword Expressions

Omid Rohanian, Shiva Taslimipoor, Victoria Yaneva and Le An Ha

Research Group in Computational Linguistics, University of Wolverhampton, UK

{omid.rohanian,shiva.taslimi,v.yaneva,ha.l.a}@wlv.ac.uk

Abstract

In recent years gaze data has been increasingly used to improve and evaluate NLP models due to the fact that it carries information about the cognitive processing of linguistic phenomena. In this paper we conduct a preliminary study towards the automatic identification of multiword expressions based on gaze features from native and non-native speakers of English. We report comparisons between a part-of-speech (POS) and frequency baseline to: i) a prediction model based solely on gaze data and ii) a combined model of gaze data, POS and frequency. In spite of the challenging nature of the task, best performance was achieved by the latter. Furthermore, we explore how the type of gaze data (from native versus non-native speakers) affects the prediction, showing that data from the two groups is discriminative to an equal degree. Finally, we show that late processing measures are more predictive than early ones, which is in line with previous research on idioms and other formulaic structures.

1 Introduction

In order to alleviate the burden that language comprehension poses on the short-term memory, the human brain uses frequently occurring formulaic sequences such as multiword expressions, collocations and idioms, among others, and stores them as units in the long-term memory (Conklin and Schmitt, 2012). As a result of the efficacy of this approach, a large proportion of the spoken and written language is formulaic, with some corpus studies claiming that between 52% and 58% of the language in the analysed corpora falls into this category (Erman and Warren, 2000), and other studies claiming that this figure is around 32% (Foster, 2001). Given the frequency with which this phenomenon occurs, the automatic identification of formulaic language is of paramount importance for a number of Natural Language Processing (NLP) tasks and applications.

Conklin and Schmitt (2012) argue that our brains store and process very frequent and highly fixed combinations as "wholes" as opposed to single words being added together and that this difference in processing is reflected in eye tracking data. A number of eye tracking studies discussed in Section 2 show that there is a processing advantage for formulaic sequences for both native and non-native speakers compared to controlled non-formulaic sequences. Based on this evidence, it could be concluded that the characteristics of formulaic language could be captured through differences in the gaze patterns between formulaic and non-formulaic sequences. In a similar way, gaze data has previously been successfully used in other NLP tasks such as part-of-speech tagging (Barrett et al., 2016a) and evaluation of word embeddings (Søgaard, 2016), and it has been shown that gaze signals transfer across languages (Barrett et al., 2016b). In this sense, automatically identifying formulaic sequences based on gaze features could not only contribute to potentially improving classification accuracy and gaining insight into the cognitive processing of such units, but can also provide a language-independent approach to identification of formulaic phrases. However, it is important to note that almost all studies using gaze data to investigate formulaic language focus solely on idioms and that other types of formulaic units have been significantly understudied.

In the present research we conduct a preliminary study towards the identification of multiword expressions (MWEs) based on gaze features.

Proceedings of Recent Advances in Natural Language Processing, pages 601–609,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_078

An MWE is commonly known as a combination of two or more words, not necessarily continuous, that pose difficulties on language processing (Sag et al., 2002) and that typically have syntactic and semantic idiosyncrasies (Fazly and Stevenson, 2006). In particular, we focus on two common types of MWEs, namely, Verb-Particle (e.g. *give up*) and Verb-Noun (e.g. *take place*) constructions.

We use the GECO corpus (Cop et al., 2016), a monolingual and bilingual corpus of the eye-tracking data from participants reading a complete novel. The use of this data allowed comparison between the gaze patterns of native and non-native English speakers, as well as a comparison of the predictive power of data obtained from these two groups. Furthermore, we explore a range of early and late measures of cognitive processing in order to determine which of the two groups of features carries more important linguistic information. In order to account for the fact that MWEs are often processed as unified structures, we used Conditional Random Fields (CRF) classifier to label sequences of words, together with a variety of early and late gaze features.

The contributions of this work are as follows:

- We explore a novel approach to MWE identification based on gaze data. We compare a POS + Frequency baseline to: i) a prediction model based solely on gaze features and ii) a prediction model based on gaze features, POS, and frequency.
- A comparison between the predictive power of gaze data from native and non-native speakers of English in the context of MWE identification.
- A comparison between the predictive power of a number of early and late gaze features in the context of our current task.

The code used in the experiments and the annotation of the MWEs are made freely available[1]. The GECO corpus could be downloaded freely at: `http://expsy.ugent.be/downloads/geco`. For an investigation discussing the cognitive processing of the MWEs in the corpus, refer to (Yaneva et al., 2017).

The rest of this paper is organised as follows. Section 2 presents related work from the fields of eye tracking and MWEs research, while Section 3 describes the data used in this study and Section 4 describes the gaze features. The experimental approach and the actual experiments conducted are presented in Section 5, and are then reported and discussed in Sections 6 and 7, respectively. Finally Section 8 contains the main conclusions and avenues for future work.

[1] `https://github.com/omidrohanian/gaze-mwe-ranlp2017`

2 Related Work

This section presents related work from the fields of eye tracking research and automatic identification of multiword units.

2.1 Eye Tracking and Formulaic Language

Eye tracking is a process where an eye-tracking device measures the point of gaze of an eye (gaze fixation) or the motion of an eye (saccade) relative to the head and a computer screen (Duchowski, 2009). Fixations are eye movements which stabilise the retina over a stationary object of interest, which, in the case of reading research, is the written text and its units (letters, words, phrases, etc). Gaze fixations and revisits (go-back fixations to a previously fixated object) have been widely used as measures of cognitive effort by taking into account their durations and the places in text where longer fixations occur (Duchowski, 2009). Early gaze measures such as first fixation duration give information about the early stages of lexical access and syntactic processing, while late gaze measures such as total dwell time or total number of fixations give information about late stages of processing (e.g. late syntactic processing, textual integration processes, lexical and syntactic/semantic processing and disambiguation in general). A series of studies on eye tracking during reading show that gaze data is sensitive to linguistic phenomena such as word frequency, verb complexity and lexical ambiguity, as well as contextual effects on word perception (Rayner, 1975; Rayner and Duffy, 1986; Rayner, 2009; Rayner et al., 2012).

Gaze data has been previously used to investigate formulaic language with a main focus on idiom research (Underwood et al., 2004; Siyanova-Chanturia et al., 2011; Conklin and Schmitt, 2012; Siyanova-Chanturia, 2013; Cutter et al., 2014; Carrol and Conklin, 2015). For example, Underwood et al. (2004) showed that native speakers read idioms faster and with fewer fixations compared to control non-idiomatic phrases and

that the last word of the idiom was read faster than the last word in the control condition. Similarly, non-native readers produced fewer fixations when reading idioms than when reading control phrases but there were no differences in the durations of those fixations (Underwood et al., 2004). Siyanova-Chanturia et al. (2011) corroborated the processing advantages of idioms over novel phrases and showed that idioms required less re-reading and less re-analysis. Interestingly, there were no significant differences in the early gaze measures, suggesting that early eye-tracking measures may not be suitable for investigation of formulaic language (Siyanova-Chanturia et al., 2011). This result may be explained with previous research on predictability of single words showing strong effects in terms of shorter first fixation durations and greater likelihood of skipping (Rayner and Well, 1996). However, Carrol and Conklin (2015) argue that this effect may not scale up to formulaic units in a simple fusion and suggest taking an approach balancing between local, lexical context and global discourse context. Assuming that the case of formulaic language is that "the whole is greater than the sum of the parts", Carrol and Conklin (2015) suggest the use of a *hybrid* approach where formulaic language is analysed both as a whole and at the level of individual words. In order to partly account for this effect we use an algorithm which represents the data as a sequence of words considering their neighbouring word features.

2.2 Identification of MWEs

MWEs have been investigated in computational linguistics based on their many different characteristics such as fixedness (Fazly and Stevenson, 2008), non-compositionality (Baldwin and Kim, 2010), and semi-productivity (Villavicencio, 2003). We have used these properties as the main guidelines for annotating MWEs, specifically following the guidelines provided by the PARSEME project on identifying verbal MWEs.[2] High frequency of MWEs and in particular, the principle that MWEs usually are constructed from high frequency word components have been studied extensively in computational linguistics (Granger and Meunier, 2008; Fazly, 2007).

In the most recent MWE workshop (Savary et al., 2017), several language-independent systems have been proposed for identifying or extracting MWEs. When used in conjunction with CRF models (Scholivet and Ramisch, 2017) or structured perceptrons (Schneider et al., 2014), Part-of-Speech (POS) tags have been shown to be useful features (especially when parsing information is not available) to identify MWEs. Schneider et al.'s (2014) statistical sequence model has achieved the best F1-score of 60% in identifying all heterogeneous types of MWEs and truly shows how challenging the task is.

3 Eye Tracking Data

The GECO corpus (Cop et al., 2016) used in this study is, to the best of our knowledge, the most recent eye tracking corpus for English, which: i) contains gaze data from a natural reading task (as opposed to e.g. single sentences), ii) is long enough to contain a sufficient number of MWEs, and iii) contains paired gaze data from native and non-native readers. Eye tracking data was collected for both the English version of the novel and its translation in Dutch; however, in the current study we only focus on the English part of the data.

The text of the corpus is a novel by Agatha Christie entitled "The Mysterious Affair at Styles", the English version of which contains 54,364 tokens and 5,012 unique types. The novel was selected based on the fact that its word frequency distribution had considerable similarity to the one in natural language use, as observed in the Subtlex database (Cop et al., 2016). The novel was read by 14 English monolingual undergraduates from the University of Southampton and 19 Dutch (L1) - English (L2) bilingual students at Ghent University (intermediate and advanced). The two groups were matched on age and education level. The monolingual participants read only the English version of the novel, which amounted to a total of 5,031 sentences. The bilingual participants read chapters 1 - 7 in one language and 8 - 13 in the other in a counterbalanced order, thus reading 2,449 English sentences. The eight bilingual participants who read the first part of the novel in English read 2,852 English sentences.

The sampling rate of the eye tracking device was 1 kHz. Full details about the method and procedure used for the development of the corpus could be found in (Cop et al., 2016).

[2] https://typo.uni-konstanz.de/PARSEME/images/shared-task/guidelines/PARSEME-ST-annotation-guidelines-v6.pdf

4 Gaze Features

A number of gaze features were selected for the corpus and are listed in Table 1. All gaze features were averaged over 14 native readers for one set and 19 non-native readers for another set of data. We divided the features into *early* and *late* processing measures. Early measures capture processes such as lexical access and syntactic processing, as well as oculomotor processes and visual properties of the region. An example of such a measure is *first fixation duration* (Demberg and Keller, 2008). Late measures account for late syntactic processing, textual integration processes, lexical and syntactic/semantic processing and disambiguation in general. An example of a late measure is the *total reading time* of a region, which is the sum of all fixations on a region, including refixations of the region after it was left (Demberg and Keller, 2008).

Table 1: Categorised Gaze Features

Early	WORD_FIRST_FIXATION_DURATION WORD_FIRST_RUN_FIXATION_COUNT WORD_FIRST_RUN_FIXATION_% WORD_FIRST_FIXATION_VISITED_WORD_COUNT WORD_FIRST_FIX_PROGRESSIVE WORD_SKIP
Late	WORD_FIXATION_COUNT WORD_FIXATION_% WORD_RUN_COUNT WORD_GO_PAST_TIME WORD_SELECTIVE_GO_PAST_TIME WORD_TOTAL_READING_TIME WORD_TOTAL_READING_TIME_% WORD_SPILLOVER WORD_AVERAGE_FIX_PUPIL_SIZE WORD_SECOND_FIXATION_DURATION WORD_SECOND_RUN_FIXATION_COUNT WORD_SECOND_RUN_FIXATION_% WORD_SECOND_FIXATION_RUN WORD_THIRD_FIXATION_DURATION WORD_THIRD_RUN_FIXATION_COUNT WORD_THIRD_RUN_FIXATION_% WORD_THIRD_FIXATION_RUN WORD_LAST_FIXATION_DURATION WORD_LAST_FIXATION_RUN

5 Experiments

This section presents the annotation procedure, method, and setup used to conduct the experiments, as well as the definition of the baseline.

5.1 Annotation

Two annotators with linguistic background labelled the GECO corpus for Verb + Noun and Verb + Particle constructions. The procedure was as follows. Both annotators read the entire corpus (as opposed to annotating automatically extracted cases) and marked both types of MWEs by considering cases where the components of an MWE can occur with at most three words in between. All Verb + Noun and Verb + Particle expressions (with or without gaps) irregardless of whether they were annotated as MWE or not are considered for evaluating the agreement between the annotators. The kappa inter-annotator agreement is k = 0.7864. Furthermore, we have resolved the annotation differences by employing a third annotator to decide in cases of disagreement.

In order to prepare sequences to be trained by the CRF model, we extract from the corpus all patterns of Verb + Noun and Verb + Prepositions (and Verb + a list of other particles such as *up, down, over, etc*) with at most three words between the components. MWEs are tagged using the IOB format based on the annotations. The (B) tag stands for words appearing at the beginning, (I) for words occurring inside, and (O) for words that are outside of an MWE (Sang, 2002). Verb + Noun and Verb + Particle patterns, with a window of one word before and one word after, are fed into the CRF model as input sequences. In total, there are 381 sequences that contain MWEs and 5,837 which do not. Two examples of annotated sequences are as follows. The first sequence contains an MWE while the second does not.

1)	*have*	*knocked*	*us*	*all*	*down*	*with*	*a*
	O	B	O	O	I	O	O

2)	*have*	*been*	*asked*	*both*	*by*	*my*
	O	O	O	O	O	O

5.2 Method

For our task of sequence labelling with sparse data, we use Conditional Random Fields (CRFs). CRFs are capable of relaxing the strong independence assumptions present in similar models like HMMs, which make them a suitable choice in a structured prediction task where context is of importance (Lafferty et al., 2001).

We use Pycrfsuite[3] which is a freely available Python wrapper around the crfsuite toolkit[4]. For the training algorithm we use Adaptive Regulari-

[3] https://python-crfsuite.readthedocs.io/en/latest/

[4] http://www.chokkan.org/software/crfsuite/

Algorithm 1 Bootstrap aggregating on CRF labels

procedure BAGGING
  $[] \leftarrow result$
  $n_{test} \leftarrow \frac{1}{5}\|MW\|$
  $n_{train} \leftarrow \frac{4}{5}\|MW\|$
  $subIter \leftarrow \frac{\|MW\| - n_{test}}{n_{train}}$
  for 100 times **do**
    $test \leftarrow (sample\, of\, size\ n_{test}\ from\, MW) \cup (sample\, of\, size\ n_{test}\ from\, nonMW)$
    for i = 1 to subIter **do**
      $train \leftarrow (sample\, of\, size\ n_{train}\ from\ (nonMW - test)) \cup (MW - test)$
      $C_i \leftarrow CRF(train, test)$
    $C^* \leftarrow [\underset{y \in Y}{argmax} \sum_{i:C_i[0]} 1, ..., \underset{y \in Y}{argmax} \sum_{i:C_i[2*n_{test}]} 1]$
    $result.add(Eval(C^*))$
  return $mean(result), std(result)$

sation Of Weight Vector (AROW) that is suitable for handling inherently noisy labels in the training set (Crammer et al., 2009).

In order to extract features for the CRF model, given each sequence:

1. gaze features of each word in the sequence are added;

2. for the verb part of the sequence, we also add the features of the last component of the pattern (Verb + Noun or Verb + Particle);

3. for all other words of the sequence, on the other hand, we add the features of the verb component of the pattern.

The gaze features of the GECO corpus, used in this study are listed in Table 1.

5.3 Setup

In order to tackle the imbalance of data, we employ a bootstrap aggregating strategy (Breiman, 1996). We first randomly select one fifth of the MWEs and the same number from non-MWEs as the test data. Then, we divide the remaining non-MWEs to several different sections with the same size as the remaining MWEs. We train the model on each section of non-MWEs and the whole training set of MWEs. We test the model on the held-out test data by obtaining the majority votes of different training models over the test sample. This process is performed 100 times and the average and standard deviations of the precision, recall and F1-score measures are reported. The formalised approach is presented in Algorithm 1.

5.4 Baseline

We apply the same CRF and aggregating approach only with lexical features as the baseline. POS and word frequency are used as the features. The GECO data is provided with the POS tags for the words, while word frequencies are derived from the BNC corpus (Leech, 1992).

In the case of these lexical features, given each word feature f_w present in the input sequence, contextual features f_{w-1} and f_{w+1} are automatically retrieved and added to the feature set. This informs the model of what is happening in the immediate neighbourhood of each word in the sequence.

6 Results

We report the results of CRF labeling using different sets of features, including POS tags, Frequency (referred to as FREQ), Early and Late gaze measures (Table 2).

Since most of the data are not MWEs and are thus irrelevant to the task, we report the results exclusively for the words at the beginning of the MWEs (B-MWE) and other words occurring within and at the end of the expressions (I-MWE).

In Table 2, we have first shown that augmenting the lexical features (POS and FREQ) with Gaze has slightly improved the performance (F = 70.05 for B-MWE and F = 54.0 for I-MWE) compared to the baseline (F = 63.6 for B-MWE and F = 48.06 for I-MWE). Although, based on the reported standard deviation measures, adding Gaze features might not be helpful in some parts of the data, in general, the combination of lexi-

Table 2: The performance (%) and Standard Deviation (std) (%) of CRF labeling models using different sets of features.

Features		Precision (std)	Recall (std)	F1-score (std)
FREQ	B-MWE	46.92 (12.17)	27.59 (13.89)	32.53 (12.03)
	I-MWE	37.00 (14.18)	10.09 (7.06)	14.76 (8.62)
POS	B-MWE	59.14 (4.75)	63.34 (11.92)	60.05 (6.46)
	I-MWE	56.43 (5.44)	39.03 (8.59)	45.44 (6.05)
POS + FREQ	B-MWE	59.95 (3.54)	68.26 (7.96)	63.6 (4.45)
	I-MWE	55.19 (4.78)	43.16 (7.77)	48.06 (5.56)
Gaze features (Early and Late)	B-MWE	51.43 (3.19)	55.55 (9.2)	53.06 (5.22)
	I-MWE	37.43 (5.95)	22.97 (6.07)	27.97 (5.19)
POS + FREQ + Gaze	B-MWE	66.68 (3.36)	74.03 (5.45)	**70.05 (3.48)**
	I-MWE	59.08 (4.8)	50.03 (5.87)	**54.0 (4.41)**
Early features	B-MWE	51.77 (5.14)	55.28 (21.74)	51.02 (12.94)
	I-MWE	37.53 (19.25)	9.73 (10.41)	13.38 (11.7)
Late features	B-MWE	50.16 (3.22)	56.06 (9.41)	52.54 (5.11)
	I-MWE	38.07 (5.0)	21.23 (6.21)	26.8 (5.84)
POS + FREQ + Early features	B-MWE	66.54 (3.68)	74.45 (6.73)	70.11 (3.82)
	I-MWE	60.01 (4.53)	49.16 (5.67)	53.85 (4.1)
POS + FREQ + Late features	B-MWE	65.0 (3.43)	74.12 (5.96)	69.59 (3.69)
	I-MWE	58.85 (4.19)	50.23 (5.77)	53.93 (3.90)

cal features and the gaze information outperforms the baseline model and the model that uses Gaze features alone (Early and Late) ($F = 53.06$ for B-MWE and $F = 27.97$ for I-MWE).

We also compare the performance of Early and Late features in identifying MWEs in the second part of the table. We note that Late features appear to be more discriminative than Early features in identifying MWEs. Although in case of the B-MWE, the improvement over Early features is minimal, the difference is more contrastive for I-MWE. Also the standard deviation for the model using Late features confirms its superior reliability. We see these improvements when using Early or Late features by themselves and not in conjuction with POS+FREQ.

Furthermore, we have conducted an experiment with gaze features extracted from non-native speakers of English. Table 3 presents a comparison between the F1 scores obtained from training on L1 and L2 gaze data. There were no significant differences between the two groups in terms of precision and recall, hence, for the purpose of brevity, we present the comparison only in terms of F1 scores. The better performance when using Late gaze features over Early is well reiterated in the data extracted for non-native speakers in this table.

Table 3: The performance (F1-score%) comparison between data from native (L1) and non-native (L2) speakers.

Features		L1	L2
Gaze	B-MWE	53.06 (5.22)	54.26 (4.8)
	I-MWE	27.97 (5.19)	26.66 (5.11)
POS + FREQ + Gaze	B-MWE	70.05 (3.48)	69.66 (3.07)
	I-MWE	54.0 (4.41)	52.84 (3.89)
Early Gaze	B-MWE	51.02 (12.94)	51.69 (13.3)
	I-MWE	13.38 (11.07)	11.63 (11.66)
Late Gaze	B-MWE	52.54 (5.11)	54.95 (5.04)
	I-MWE	26.8 (5.84)	27.24 (5.78)

7 Discussion

In this section we discuss the results presented above with regards to: i) MWEs identification accuracy, ii) comparison between the predictive power of gaze data of native versus non-native speakers, and iii) the predictive power of Early versus Late gaze features.

In terms of identification accuracy for MWEs, best performance was achieved by the model combining POS + Frequency + Gaze data for both the beginning of the MWE ($F = 70.05$), and the words occurring inside the MWE ($F = 54.0$). Even though gaze features on their own performed significantly worse than the baseline, the combined

model of Gaze + Freq + POS outperformed the baseline and achieved a performance comparable to the state-of-the-art in the field (Section 2.2). The lower values for the standard deviations in the combined model for both B-MWE and I-MWE also show that it is more reliable than the baseline in its prediction over 100 iterations. Furthermore, the fact that gaze features improve the classification accuracy means that readers process these structures differently.

We do not observe significant differences in model accuracy when running parallel models on the data from the native speakers and the one from the non-native speakers, which indicates that both data sets are discriminative to an equal extent. It is important to note that the non-native speakers were highly proficient in English and that this result may not be replicated with gaze data from less proficient readers. From a practical perspective this is important with regards to the type of eye-tracking corpora which could be used in similar experiments in the future. Since such resources are scarce and expensive to obtain, it is reassuring to know that data from non-native speakers could be used equally well for the purpose of automatically identifying MWEs. From a psycholinguistic perspective however, this finding is not in line with previous research on the differences in gaze patterns between native and non-native speakers reading formulaic language (Section 2.1). One reason for this could be that previous research using gaze data to explore the processing of formulaic language has focused predominantly on idioms, while we discuss MWEs. Another reason for this could be the different data sets used in these studies and conclusive results can only be drawn if idiom research is performed using the GECO corpus or vice-versa.

Finally, much in line with previous studies (e.g. Siyanova-Chanturia (2013)) we observe that early gaze features are not useful metrics for investigating formulaic language. It is important to note that late features were more discriminative even without using the entire Late feature set; there were no significant differences in performance when removing late features related to the third run and last runs ($F = 0.52$ for B-MWE and $F = 0.23$ for I-MWE). In our experiments the late features were particularly better at identifying the words inside the MWEs and we hypothesise that this effect could be due to the fact that given our pattern of Verb + Noun and Verb + Particle constructions, these were the disambiguation regions of the MWEs. Another possible explanation for the superiority of late features could be that mental processing of MWEs occurs after the fact, meaning, after the word is first encountered in reading. Therefore, early gaze features are not expected to contain much information about whether a particular sequence of tokens are MWE or not.

Some of the limitations of this research are related to averaging of data from multiple participants and the fact that the newly-released GECO corpus (Cop et al., 2016) has not yet been studied in detail and thus it is possible that it contains inaccuracies yet to be spotted. We plan to address the first limitation by conducting a study where separate models are built for each individual participant. This would allow analysis of individual differences and the effects they have on the robustness of the model. We chose to use the GECO corpus since it was the only corpus available which allowed comparison of gaze data from native versus non-native speakers. Nevertheless, it would be interesting to compare our current results on the GECO data to results from more established eye tracking corpora such as the Dundee corpus (Kennedy et al., 2013) in order to further assess the validity of our findings.

8 Conclusions

This paper presents preliminary research towards using gaze data to automatically identify multiword expressions. We show that MWEs are indeed viewed differently and that best classification performance is achieved by a combined model of gaze features, frequency and POS tags, which outperform models based on frequency and POS only and on gaze features only. Furthermore, we show that there is no statistically significant difference between the performance of models using gaze data from native versus highly proficient non-native speakers of English, suggesting that data from both reader groups could be used for similar tasks in the future. Finally, consistent with previous research in the field, we show that late gaze features are better predictors of formulaic language.

Future work includes incorporating different sequence labeling models (including at the level of individual participants) and replicating the experiment with gaze data from different corpora.

References

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In *Handbook of Natural Language Processing, second edition.*, CRC Press, pages 267–292.

Maria Barrett, Joachim Bingel, Frank Keller, and Anders Søgaard. 2016a. Weakly supervised part-of-speech tagging using eye-tracking data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. volume 2, pages 579–584.

Maria Barrett, Frank Keller, and Anders Søgaard. 2016b. Cross-lingual transfer of correlations between parts of speech and gaze features. In *26th International Conference on Computational Linguistics (coling)*.

Leo Breiman. 1996. Bagging predictors. *Machine learning* 24(2):123–140.

Gareth Carrol and Kathy Conklin. 2015. Eye-tracking multi-word units: some methodological questions. *Journal of Eye Movement Research* 7(5).

Kathy Conklin and Norbert Schmitt. 2012. The processing of formulaic language. *Annual Review of Applied Linguistics* 32:45–61.

Uschi Cop, Nicolas Dirix, Denis Drieghe, and Wouter Duyck. 2016. Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior research methods* pages 1–14.

Koby Crammer, Alex Kulesza, and Mark Dredze. 2009. Adaptive regularization of weight vectors. In *Advances in neural information processing systems*. pages 414–422.

Michael G. Cutter, Denis Drieghe, and Simon Liversedge. 2014. Preview benefit in english spaced compounds. *Experimental Psychology Learning Memory and Cognition* 40(6).

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109(2):193–210.

Andrew Duchowski. 2009. *Eye Tracking Methodology: Theory and Practice*. Springer, second edition.

Britt Erman and Beatrice Warren. 2000. The idiom principle and the open choice principle. *Text-Interdisciplinary Journal for the Study of Discourse* 20(1):29–62.

Afsaneh Fazly. 2007. *Automatic Acquisition of Lexical Knowledge about Multiword Predicates*. Ph.D. thesis, Department of Computer Science, University of Toronto.

Afsaneh Fazly and Suzanne Stevenson. 2006. Automatically constructing a lexicon of verb phrase idiomatic combinations. In *In Proceedings of EACL-06*. pages 337–344.

Afsaneh Fazly and Suzanne Stevenson. 2008. A distributional account of the semantics of multiword expressions. *Italian Journal of Linguistics* 1(20):157–179.

Pauline Foster. 2001. Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. *Researching pedagogic tasks: Second language learning, teaching, and testing* pages 75–93.

Sylviane Granger and Fanny Meunier. 2008. *Phraseology: an interdisciplinary perspective*. John Benjamins Publishing Company.

Alan Kennedy, Joël Pynte, Wayne S Murray, and Shirley-Anne Paul. 2013. Frequency and predictability effects in the dundee corpus: An eye movement analysis. *The Quarterly Journal of Experimental Psychology* 66(3):601–618.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*. volume 1, pages 282–289.

Geoffrey Leech. 1992. 100 million words of english: the british national corpus (bnc). *Language Research* 28(1):1–13.

Keith Rayner. 1975. The perceptual span and peripheral cues in reading. *Cognitive Psychology* 7(1):65–81.

Keith Rayner. 2009. Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology* 62(8):1457–1506.

Keith Rayner and Susan A Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition* 14(3):191–201.

Keith Rayner, Alexander Pollatsek, Jane Ashby, and Charles Clifton Jr. 2012. *Psychology of reading*. Psychology Press.

Keith Rayner and Arnold D Well. 1996. Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review* 3(4):504–509.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann A. Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*. Springer-Verlag, London, UK, UK, CICLing '02, pages 1–15.

EF Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition .

Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang Qasemizadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, et al. 2017. The parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. pages 31–47.

Nathan Schneider, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014. Discriminative lexical semantic segmentation with gaps: Running the mwe gamut. *Transactions of the Association for Computational Linguistics* 2:193–206.

Manon Scholivet and Carlos Ramisch. 2017. Identification of ambiguous multiword expressions using sequence models and lexical resources. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Association for Computational Linguistics, Valencia, Spain, pages 167–175.

Anna Siyanova-Chanturia. 2013. Eye-tracking and erps in multi-word expression research: A state-of-the-art review of the method and findings. *The Mental Lexicon* 8(2):245–268.

Anna Siyanova-Chanturia, Kathy Conklin, and Norbert Schmitt. 2011. Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research* 27(2):251–272.

Anders Søgaard. 2016. Evaluating word embeddings with fmri and eye-tracking. *ACL 2016* page 116.

Geoffrey Underwood, Norbert Schmitt, and Adam Galpin. 2004. The eyes have it. *Formulaic sequences: Acquisition, processing, and use* 9:153.

Aline Villavicencio. 2003. Verb-particle constructions and lexical resources. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*. Association for Computational Linguistics, Stroudsburg, PA, USA, MWE '03, pages 57–64.

Victoria Yaneva, Shiva Taslimipoor, Omid Rohanian, and Le An Ha. 2017. Cognitive processing of multiword expressions in native and non-native speakers of english: Evidence from gaze data. In *Mitkov, R. (Ed.) Computational and Corpus-based Phraseology (to be appeared)*. Springer: Heidelberg, New York, London.